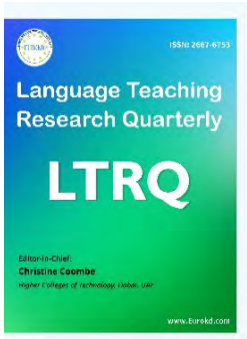

Language Teaching Research Quarterly

2023, Vol. 37, 161–178

The Influence of Passage Cohesion on Cloze Test Item Difficulty

Jonathan Trace

Keio University, Japan

Received *11 April 2023* ***Accepted*** *22 September 2023*

Abstract

The role of context in cloze tests has long been seen as both a benefit as well as a complication in their usefulness as a measure of second language comprehension (Brown, 2013). Passage cohesion, in particular, would seem to have a relevant and important effect on the degree to which cloze items function and the interpretability of performances (Brown, 1983; Dastjerdi & Talebinezhad, 2006; Oller & Jonz, 1994). With recent evidence showing that cloze items can require examinees to access information at both the sentence and passage level (Trace, 2020), it's worthwhile to now look back and examine the relationship between aspects of passage cohesion—referential cohesion, semantic overlap, and incidence of conjunctives—and item difficulty by classification. The current study draws upon a large pool of cloze test passages and items (k = 377) originally used by Brown (1993) along with automated text analysis of cohesion (*Coh-Metrix*, McNamara et al., 2014) to examine the impact of passage cohesion on item function. Correlations, factor analysis, and linear regression point to clear though minimal differences for both sentential and intersentential items as they relate to aspects of passage cohesion, the results of which may inform future test design and interpretation of cloze performance.

Keywords: *Cloze Tests, Cohesion, Text Analysis, Validity*

Introduction

Few researchers have contributed more to our understanding on the function, use, and interpretability of cloze tests than James Dean (JD) Brown. Ten years ago, JD himself reflected upon his own twenty-five-year legacy of studying cloze tests by stating that, "from the outset, I was fascinated by cloze tests because I believed that they function well as overall ESL proficiency tests even though we have very little idea how they work" (2013, p. 2). And it is from this initial point of curiosity that he set out to explore and discover much about what we currently know about cloze tests and their many uses. Indeed, there appear to be few aspects of their function and design that remain untouched in some way by his always careful, always honest, and always engaging research into these simple yet puzzling tools, from scoring

* Corresponding author.
E-mail address: tracej@sfc.keio.ac.jp
https://doi.org/10.32038/ltrq.2023.37.08

approaches (Brown, 1980) to reliability and validity (Brown, 1984; Brown et al., 2012), item discrimination (Brown, 1988; Brown et al., 2001; Brown et al., 2016), cohesion (Brown, 1983; 1989), passage qualities (Trace et al., 2017) and even their use outside of language assessment (Brown & Grüter, 2022).

One question addressed by Brown in 1993 and taken up later by Trace (2016; 2020) is the degree to which cloze items can and do draw upon contextual information beyond the sentence level. Given that one of the benefits to using cloze tests is the context that is provided around each item, it is expected that they can measure comprehension beyond the local (i.e., sentential) syntactic structure, but also at the argument (i.e., intersentential) level as well. As this has implications for construct validity (Trace, 2020), not to mention passage and item selection, a fuller view of the impact of passage-level features is required for a better understanding of how these assessments work and what they can tell us about test-taker performance.

The current study follows up on earlier works by JD by taking into consideration the effects and qualities of cloze passages from which items are drawn, examining what, if any, factors of cohesion are contributing to systematic differences in item function. It is hoped that through this understanding of passage influence, a more complete and informed approach to the construction and interpretation of cloze tests can be put into practice in the future.

Literature Review

Cloze Design

Originally developed by Taylor (1953), cloze tests were intended to measure the comprehensibility of textbooks in English, before being later adapted as a measure of L1 reading ability (cf. Bormuth, 1967) and eventually applied to L2 contexts as well (cf. Alderson, 1979). While debate persists about the degree to which cloze tests function similarly or differently across the two contexts (Gellert & Elbro, 2013), recent work by Trace (2020) has pointed to similarities in performance across both L1 and L2 examinees indicating that cloze tests may be measuring along similar constructs of reading ability, though it may also be criterion-valid as an indirect measure of general language proficiency (cf. Bachman, 1985). As with anything, function depends upon use and the context in which tests are administered and interpreted.

Putting questions of validity aside for the moment, one of the endearing—as well as frustrating—qualities of cloze tests that make them different from other discrete measures of ability is that the items are placed within a rich contextual field, namely the passage from which they are drawn. The longstanding assumption with cloze test items is that because they are taken from a larger text, that same text will be essential when it comes to producing answers. In other words, successfully reconstructing the originally intended meaning of the passage requires that test-takers will draw upon information that is both locally obtained within the immediate grammar and context of the passage, as well as more global or thematic information found at the broader levels of cohesion and argument structure.

This argument has been framed by JD and others (Abraham & Chapelle, 1992; Bachman, 1985; Jonz, 1990) as a question of whether cloze items rely upon context at the sentential level alone, or if they can also access information at the intersentential level (i.e., cross-paragraph, composite information). While earlier studies came out in favor of cloze items tapping primarily—if not exclusively—into sentential information (Shanahan et al., 1982), more recent

work on the topic by Brown and others (Brown et al., 2016; Kleijn et al., 2019; Trace et al., 2017) has found increasing evidence that items can draw upon information at the intersentential level.

Recently, Trace (2016, 2020) used large-scale L1 data gathered using *Mechanical Turk* to classify items by the amount of context required to consistently reproduce a correct answer on a series of cloze tests and items. Through the manipulation of contextual clues, he gathered clozentropy scores (i.e., the likelihood in which an L1 user of English was able to produce a correct answer) to categorize items as sentential or intersentential. The findings indicated that not only were both item types present in relatively equal number, but that intersentential items were not more difficult on average than other types of items found within the tests, a claim earlier made by Kobayashi (2002).

Passage Cohesion

The relationship between context and item difficulty has often been tied to notions of cohesion (e.g., Chihara et al., 1994; Dastjerdi & Talebinezhad, 2006; Oller & Jonz, 1994). Passages that contain more references, overlap semantically, and express connections between themes and ideas through conjunctives are thought to be more internally cohesive (Eggins, 2004; Halliday & Hasan, 1976), which can help with readers' comprehension of the text (McNamara et al., 2014).

Studies in both L1 and L2 research have found that cohesive devices (i.e., references, lexical synonyms, and conjunctions) can and do function as items in cloze tests (e.g., Brown, 1983; Gellert & Elbro, 2013), though less evidence exists that these draw upon context at the intersentential level equally. While references and conjunctions both seem to require passage-level comprehension by examinees, items that draw upon semantic relationships have been found to be more locally accessed (Bridge & Winograd, 1982; Storey, 1997), and these effects may further vary by learner ability (Brown, 1989) or conjunction type (Goldman & Murray, 1992). Later work by Trace (2016) found similar results when looking at the contributions of passage cohesion on item difficulty using structural equation modeling. While cohesion as a whole contributed very little to item difficulty, conjunctions were slightly more impactful compared to references or lexical overlap, with only the latter appearing to be more difficult on average for examinees.

Research Questions

As intersentential and sentential items have the appearance of functioning similarly in terms of difficulty, the question that remains is to what degree they are influenced differently by passage-level factors of cohesion. In other words, are these items drawing upon specific types of cohesive devices within the text, and, if so, to what degree are these influencing item difficulty? To that end, the following research questions were posed:

RQ_1: What, if any, relationships are there between aspects of passage cohesion and cloze item difficulty, and are these similar for items drawing upon different levels of context?

RQ_2: To what degree do measures of cohesion form distinct factors for cloze test passages?

RQ_3: To what degree can cohesion predict cloze item difficulty, and are these predictions similar for items drawing upon different levels of context?

Methodology

Cloze Items

Data for the current study are drawn from an initial pool of 449 cloze items taken from 15, 30-item[1] cloze tests originally developed by Brown (1993). These 15 tests come from a larger battery of 50 cloze passages designed by Brown to represent generalized, written English. Due to a high number of the original 50 tests underperforming in terms of item function (i.e., difficulty and discrimination), score reliability, and score variation, a limited subset was selected for use in later studies by Trace (2016, 2020). The selection criteria were based upon those tests with a high ratio of functioning items, reliability estimates of .80 or higher, and tests displaying not markedly non-normal score distributions. Twenty-three tests were found to meet these criteria, and thus to preserve Brown's original intention of examining a generalized subset of cloze tests, 15 tests were randomly selected for analysis.

Included in this set of 449 items, 100 items were identified as drawing upon sentential (i.e., local) context only, 72 drew upon intersentential information, with the remaining items classified as either extra-textual or open-classed such that multiple answers were possible and therefore their function was unable to be determined given the scoring method employed. These items and their respective tests formed the basis for the current study.

Participants

Participant data come from previously collected test-taker data originally featured by Brown (1993) and colleagues (Brown et al., 2012). These data consist of 2,246 English as a foreign language learners from universities across Russia (n = 1548) and Japan (n = 698). As was the case in the original studies, limited participant data is available with the exception that ages ranged between 14-45 for the Russia sample and 18-24 for the Japan sample.

Instruments

Brown (1993) initially created the 50 tests by randomly selecting from books found in a North American public library and extracting 350–500-word passages. The original passages were designed to contain a clear starting point to maintain internal cohesion and include works from both fiction and non-fiction. Items in the original 50 cloze tests were created using a pseudo-random deletion approach, where every 12th word was deleted until a total of 30 items were generated per test, with the first and last sentences of each passage not containing any items to provide examinees with adequate contextual information. Because of the large-scale nature of the original studies, scoring was carried out using an exact-answer approach, meaning that only answers that replicated the originally omitted word were marked as correct. For more on test design and selection see Brown (1993) and Trace (2016).

Procedures

The original 50 tests were administered to intact English classes in 18 Japanese and 38 Russian universities. Tests were randomly distributed such that every participant had an equal chance

[1] One of the tests contained only 29 items due to an error in its original construction, so the total of 449 reflects this missing item rather than the expected total of 450.

of receiving any of the test versions. Participants had 25 minutes to complete both their assigned test as well as a small, 10-item anchor passage also used in Brown (1993).

Measures

Item function was gauged in terms of logit difficulty with Many-faceted Rasch Measurement (MFRM) using *FACETS* (Linacre, 2010). As part of the original data collection process, all participants also completed the same 10-item anchor passage. By anchoring performance, the 15 selected tests and their accompanying data could be placed upon a single logit scale with separate facets for tests, items, and participants to allow for direct comparisons in terms of item difficulty and examinee ability, following similar procedures used by Brown and colleagues (2016; Trace et al., 2017) with the full 50 tests.

In addition to item function, the 15 passages were also analyzed in terms of their cohesive features using *Coh-Metrix* 3.0 (McNamara et al., 2014). This tool was developed to provide quantitative descriptions of cohesive and syntactic qualities of written texts using 108 different indices covering a variety of categories including text length, syntactic complexity, readability, and cohesion. As the current study was primarily focused on aspects of text cohesion alone, three of the 11 categories of indices were included in the analysis. These were: (a) 12 indices for Referential Cohesion (i.e., anaphor overlap); (b) eight indices for Semantic Overlap employing latent semantic analysis; and (c) eight indices for Connectives (i.e., the presence of conjunctions) for a total of 28 separate measures. See Appendix A for a full list of indices and their descriptions.

Unlike item difficulty, passage variables contain an inherent limitation in that they are whole-passage descriptions and therefore do not display variation at the item level, which can have drawbacks for statistical analysis (Trace, 2016). To account for this, the current study employed a unique approach to coding and analyzing passage data. Rather than use the same index values for all 30 items in a single test, separate versions were created by removing the target item only, which were then run through *Coh-Metrix* individually to capture the minimal effects of the missing item's absence on passage cohesion. In other words, passage-level data for Test 1, Item 1 was gathered by analyzing the Test 1 passage with only the target word for Item 1 removed, with similar procedures carried out for the remaining 29 items. While this did not have a markedly observable impact on variation in passage-level descriptions, it did result in slightly more unique, passage-level data for each individual item in the analysis.

Analysis

In order to analyze the potential relationship between passage cohesion and item difficulty, correlations were used to compare items by difficulty to the selected cohesive indices. Indices were also examined using principal components analysis to determine whether or not they formed distinct factors in the data and gauge the degree to which passage features could reflect item variation. Lastly, multiple linear regression was used to examine which measures of cohesion could function as reliable predictors of item difficulty across different categories of item types.

Results

Data Cleaning

Item Data: Prior to analysis, the 449 cloze items were examined in terms of their function and model fit according to the assumptions of Rasch measurement. Initial MFRM analyses revealed that 19 of the items were not answered correctly by any of the examinees and given that Rasch uses a probability model to assign difficulty values, items without any correct answers can be considered outliers within the data and were therefore removed. Furthermore, 24 items were identified as misfitting the model. Eighteen of these were classified as underfitting in that their difficulty scores could not be accurately predicted due to unexplained variance (e.g., guessing on the part of test-takers). As underfitting items can limit the overall predictability of the model (Bond & Fox, 2007), these 24 items were also removed from the analysis. Lastly, six items were identified as overfitting the model, which can indicate that difficulty scores are somehow being predicted too exactly (i.e., with little to no variance). Given the small number of overfitting items, and the fact that these do not necessarily detract from the model itself, these items were retained for a total of 37 removed items.

Passage Data: Passage data were also examined in terms of normality, outliers, and collinearity. While descriptive data for the 28 indices revealed some signs of non-normality, only three variables displayed markedly non-normal distributions (CRFNO1, CRFNNOa, & LSASS1) based on skewness statistics. Because of the large sample size, it is estimated that even in cases of non-normality the sampling distributions are likely to form normal distributions and so it is unlikely that violations of normality will have an impact on later analyses. Most of the indices also revealed moderate to high standard deviations relative to their means, with no evidence of univariate outliers within the data. The presence of multivariate outliers was checked using Mahalanobis distance ($\chi^2(28) = 56.89$, $p < .0001$), with 35 items exceeding this threshold. These items were subsequently removed from the analysis, for a final total of 377 items.

Given the number of indices *Coh-Metrix* provides on textual cohesion, the potential for redundancy across measures was high, and therefore collinearity across the 28 indices was checked using Pearson correlations. Using a threshold of values .90 and higher to indicate collinearity, several instances of overlap were observed in the data, typically between indices that shared a similar category. Five indices (CRFNO1, CRFSOa, CRFCWOa, LSASS1, and CNCNeg) were removed from the dataset, eliminating all instances of collinearity for a revised total of 23 indices.

Item Descriptives

Descriptive data for the 377 items are given in Table 1, with separate groupings for each sentential, intersentential, and extra-textual items, as well as the combined dataset. Values are based on logit measures, in which lower (i.e., negative) values indicate a lack of difficulty while higher (i.e., positive) values indicate increased difficulty relative to examinee ability. The number of items per classification is given in the second column, followed by the logit average (*M*), standard deviation (*SD*), standard error (*SE*), and minimum and maximum observed values. As is indicated in the table, there were far more extra-textual items, and as stated here and elsewhere (Trace, 2020), this is primarily due to the exact-answer nature of how scores were assigned. While intersentential items were shown to be slightly more difficult than

sentential items, an independent samples *t*-test revealed that this difference was not statistically significant ($t(147) = 1.62$, $p = .11$). Extra-textual items were overall quite difficult, as evidenced by both an average logit value of 2.19 and a higher minimum of -1.22. For full summary test statistics, including reliability, see Trace (2020).

Table 1

Item Logit Descriptive Statistics

	n	*M*	*SD*	*SE*	Min	Max
Sentential	88	0.11	1.47	0.16	-3.14	4.65
Intersentential	61	0.54	1.75	0.22	-2.55	5.56
Extra-Textual	228	2.19	1.59	0.11	-1.22	5.64
Combined	377	1.44	1.85	0.10	-3.14	5.64

Passage Descriptives

Passage cohesion based on the 23 indices for textual cohesion were obtained by running each passage through *Coh-Metrix* 3.0. Table 2 provides descriptive results for each, with the index given in the first column, followed by the mean, *SD*, *SE*, minimum and maximum, and the skewness for each distribution. Indices are ordered by category, beginning with those for referential cohesion, followed by semantic overlap and connectives. Note that each index uses a slightly different scale, with measures for referential cohesion describing the average number of sentences with overlap, while semantic overlap is described along a ratio of 0.00 (i.e., no cohesion) to 1.00 (i.e., high cohesion). Connectives, likewise, are based upon incidence scores (i.e., occurrences per 1000 words). While direct comparisons are difficult, they are similarly unnecessary as the focus is on how these different indices vary in accordance to or in contrast with one another relative to item classification. Nevertheless, some trends can be observed within the data, most notably that distributions on the whole appear to be skewed (*SE* of skewness = .13), though none markedly so. Given the sample size, it is expected that violations should not impact later interpretations. More interestingly, there appears to be variation within each index, which may reflect that each was calculated at an item-level rather than a static passage-level. Indeed, when indices were examined passage by passage, slight variations were observed.

Table 2

Descriptive Data for Indices of Cohesion for Test Passages

	M	*SD*	*SE*	Min	Max	Skewness
Referential Cohesion						
CRFAO1	0.482	0.182	0.009	0.194	0.909	0.436
CRFSO1	0.321	0.227	0.012	0.077	0.833	0.831
CRFNOa	0.205	0.158	0.008	0.026	0.611	0.983
CRFAOa	0.402	0.176	0.009	0.113	0.743	-0.080
CRFCWO1	0.104	0.048	0.002	0.031	0.209	0.637
CRFCWO1d	0.117	0.031	0.002	0.063	0.169	-0.061
CRFCWOad	0.098	0.021	0.001	0.064	0.134	0.122
CRFANP1	0.414	0.184	0.009	0.150	0.686	0.016
CRFANPa	0.173	0.117	0.006	0.028	0.456	0.831
Semantic Overlap						
LSASS1d	.168	.047	.002	.094	.287	0.634
LSASSp	.187	.098	.005	.053	.429	0.652
LSASSpd	.165	.051	.003	.068	.296	0.587
LSAPP1	.309	.162	.008	.088	.618	0.236
LSAPP1d	.145	.054	.003	.037	.268	0.385
LSAGN	.330	.066	.003	.236	.506	0.954
LSAGNd	.124	.023	.001	.084	.187	0.845
Connectives						
CNCAll	95.866	18.176	0.936	65.854	129.870	0.514
CNCCaus	21.925	8.649	0.445	7.282	35.545	-0.016
CNCLogic	41.053	11.421	0.588	21.327	61.611	-0.033
CNCADC	20.129	8.890	0.458	2.667	43.902	0.538
CNCTemp	24.183	8.969	0.462	7.109	38.961	-0.208
CNCAdd	51.671	11.008	0.567	26.829	75.325	-0.172
CNCPos	81.922	15.068	0.776	56.872	106.952	0.155

Correlations

In order to gauge the relationship between passage cohesion and item difficulty, Pearson product correlations were run for each the sentential, intersentential, and extra-textual item sets across all 23 indices in comparison to logit difficulty scores (Table 3). Across the board, only weak correlations were observed for all the indices and logit difficulty, with only a few exceeding values of .100. This was to be expected, as these are still passage-level descriptions applied to item-level data, even with the analysis attempting to optimize variation.

Table 3

Pearson Correlations for Cohesion and Item Difficulty by Classification

Index	Sentential	Intersentential	Extra-Textual
CRFAO1	-.055	-.054	-.065
CRFSO1	.022	-.085	**-.115**
CRFNOa	-.003	-.043	-.092
CRFAOa	**-.156**	-.089	**-.113**
CRFCWO1	-.073	.071	-.092
CRFCWO1d	-.060	**.253**	-.075
CRFCWOad	**-.164**	-.010	**-.107**
CRFANP1	**-.190**	.046	.002
CRFANPa	**-.119**	**.197**	.052
LSASS1d	-.024	**-.125**	**-.121**
LSASSp	-.086	**-.197**	**-.136**
LSASSpd	-.062	**-.173**	**-.122**
LSAPP1	.005	-.043	**-.106**
LSAPP1d	**-.185**	-.055	**-.128**
LSAGN	-.075	**-.174**	-.094
LSAGNd	.069	-.093	-.044
CNCAll	.021	-.052	.004
CNCCaus	.077	.006	-.024
CNCLogic	.080	**.198**	-.052
CNCADC	-.054	-.070	-.023
CNCTemp	**-.239**	**-.159**	**-.188**
CNCAdd	**.118**	-.030	**.150**
CNCPos	.063	-.031	.022

Note: Correlations above +/- .100 are flagged in bold.

It is worth noting that differences were observed between relationships for logit difficulties and cohesion across all three sets of items. Regarding referential cohesion, variation in word overlap (CRFCWO1d) showed a weak but also positive relationship with intersentential items, meaning that as text variation increased, so did item difficulty. However, indices for anaphor (i.e., noun/pronoun) overlap were negatively correlated with sentential item difficulty at both the local (CRFANP1) and passage (CRFANPa) level, indicating that as the number of references pointing back to specific places in the passage increased, the difficulty for items drawing on sentence-level information slightly decreased in turn. This makes sense in that a higher number of references would increase the internal cohesion of a passage. Curiously, the opposite was found for intersentential items and anaphor overlap at the passage level (CRFANPa), with an increased presence of references indicating a slight increase in difficulty. It's possible that because intersentential items already require examinees to draw upon information at the passage level, more references would require them to connect multiple pieces of the passage at once, further increasing complexity.

For indices related to semantic overlap, the majority (LSASS1d, LSASSp, LSASSpd, LASGN) were weakly related to intersentential difficulty while also showing almost no relationship with sentential difficulty, which could be taken to mean that as overlap increases, items that rely on global or passage-level information become slightly easier, which makes sense logically. This same logic would apply to extra-textual items, except that in this case difficulty is also heavily dependent upon unobservable factors as well. Only LSAPP1d (i.e., the *SD* of overlap between adjacent paragraphs) shared a relationship with sentential items, which would indicate that as paragraphs become more varied, items relying more on local information are easier to answer. One way to interpret this is that as examinees were less able to use contextual clues within the passage itself, it may have drawn their attention to more local, and in this case relevant, clues instead.

Correlations for connective indices showed fewer differences overall across categories. Intersentential items are perhaps impacted by conjunctions affecting the internal logic of a passage (CNCLogic), while sentential items are in turn impacted by additive conjunctions (CNCAdd). As extra-textual items remain ambiguous in terms of their difficulty, it is expected that passage-features should play a minimal role in regard to their difficulty, which is reflected by the fact that correlations were weaker here for cohesive indices than for other item classifications.

Factor Analysis

In order to examine the degree to which measures of cohesion formed distinct factors relative to the cloze passages, a principal components analysis (PCA) was run for the 23 measured indices. While exploratory factor analysis (EFA) was preferred as a way to identify latent underlying factors, an initial analysis resulted in communalities greater than 1.00 (i.e., Heywood cases), which can be the result of either a sample size that is too small, weak factor loadings, or an effect of overextraction (i.e., estimating too many factors) (Cooperman & Waller, 2022). Unlike EFA, PCA includes error and unique variance, though it is also designed more for data reduction than factor identification (Tabachnick & Fidell, 2012).

As indices were drawn from three categories of cohesion, a three-component solution was exacted using varimax rotation, which revealed moderate complexity, with two measures loading across all three components (LSAGN and LSAPP1). Removing these indices resulted in a cleaner model with greatly reduced, albeit imperfect, complexity. A three-component solution was further warranted both by components reporting eigenvalues of greater than 1.00, with the lowest equaling 4.726, and scree-plot leveling effects after the third component.

Loadings are given in Table 4 and ordered by component scores starting with indices loading strongly onto the first component, followed by those loading onto the second and third components respectively. Communalities are given in the rightmost column, with the percentage of explained variance for each factor provided along the bottom row. Loadings over .350 were used to identify which variables loaded on which components as this both indicates a moderate correlation between individual measures and their respective components, as well as reflected the lowest observed loading of .358 (CNCTemp). Note that two indices were still found to be complex (CRFSO1 & LSASSp) in that they loaded across both components one and three, but for the most part a clear component structure can be observed in the data that explains around 67% of the total variance.

Table 4

Principle Component Analysis for Indices of Passage Cohesion

	C1	C2	C3	h^2
CRFCWO1	**.952**	.193	.051	.946
CRFAOa	**.877**	.337	-.137	.901
CRFAO1	**.820**	.131	.010	.690
CRFCWOad	**.817**	.218	-.103	.726
CRFNOa	**.814**	.297	.312	.848
CRFCWO1d	**.638**	-.064	.049	.414
CRFSO1	**.623**	.100	**.454**	.604
LSAPP1d	**.611**	-.046	.194	.413
CNCAll	.119	**.963**	.063	.945
CNCAdd	-.061	**.833**	.060	.701
CNCPos	.271	**.797**	.104	.719
CNCLogic	.211	**.752**	.004	.610
CNCCaus	.301	**.668**	-.115	.550
CNCADC	-.066	**.514**	-.051	.271
CNCTemp	.213	**.358**	.243	.233
LSASS1d	.209	.071	**.925**	.904
LSASSpd	.205	.090	**.842**	.759
CRFANP1	.302	.218	**-.795**	.771
CRFANPa	.148	.198	**-.793**	.690
LSASSp	**.562**	.339	**.652**	.856
LSAGNd	.274	.334	**.635**	.590
% of Explained Variance	27.197	20.505	19.629	67.331

Indices aligned mostly according to expectations, with clear components for each *referential cohesion* (C1), *connectives* (C2), and *semantic overlap* (C3). One exception was that the index for variance in semantic overlap in adjacent paragraphs (LSAPP1D) loaded together with referential cohesion, while indices for anaphor overlap (CRFANP1 & CRFANPA) loaded together with indices for semantic overlap. Otherwise, it appears evident that this particular subset of measured indices represents three unique levels of cohesion across the cloze passages.

Linear Regression

Lastly, in order to determine the degree to which passage factors can predict or potentially influence item difficulty, linear regression was run for items by classification. Given the number of indices and the fact that different relationships were observed relative to item difficulty and classification, stepwise regression was used to select and optimize only those measures with sufficient predictive power. Three separate multiple linear regressions were run with logit difficulty as the predicted variable, the results of which are given in Tables 5 through 7 below (sentential, intersentential, and extra-textual items respectively). Regression

coefficients (*B*) are given in the second column, followed by the *SE*, critical-*t* values, and significance, with the total explained variance given in the final column (R^2).

Based on the findings, only one index for conjunctives (CNCTemp) was found to be a significant predictor of sentential item difficulty (p = .025), explaining only ~6% of the total variance. The model for intersentential items was found to be more interesting, with one index for each category of cohesion functioning as a significant predictor, CRFCWO1d (p = .001), LSASSp (p = .001), and CNCLogic (p = .022). Combined, these accounted for around 27% of the variance for item difficulty. While extra-textual item difficulty was also predicted by three indices, these were limited to conjunctions and referential cohesion alone, and altogether only accounted for around 9% of the total variance.

Table 5

Summary Statistics for Stepwise Linear Regression of Predictors for Sentential Difficulty

	B	*SE*	*t*	*p*	R^2
Intercept	1.08	0.452	2.390	.019	
CNCTemp	-0.038	0.017	-2.284	.025	
					.057

Table 6

Summary Statistics for Stepwise Linear Regression of Predictors for Intersentential Difficulty

	B	*SE*	*t*	*p*	R^2
Intercept	-2.681	1.029	-2.607	.012	
CRFCWO1d	24.180	7.238	3.341	.001	
LSASSp	-8.146	2.159	-3.774	< .001	
CNCLogic	0.047	0.020	2.362	.022	
					.268

Table 7

Summary Statistics for Stepwise Linear Regression of Predictors for Extra-Textual Difficulty

	B	*SE*	*t*	*p*	R^2
Constant	1.880	0.526	3.576	< .001	
CNCTemp	-0.038	0.012	-3.178	.002	
CNCAdd	0.033	0.010	3.365	< .001	
CRFAOa	-1.288	0.604	-2.131	.034	
					.089

Discussion

RQ_1: What, if any, relationships are there between aspects of passage cohesion and cloze item difficulty and are these similar for items drawing upon different levels of context?

Findings indicate that relationships between passage cohesion and item difficulty are minimal at best, with correlations for each sentential, intersentential, and extra-textual items found to be quite weak overall. Most were also found to be inversely related to item difficulty (i.e.,

negatively correlated), indicating that as textual cohesion decreases, item difficulty increases on a small scale.

Interestingly, aspects of cohesion seem to correlate differently depending on item classification. Indices for referential cohesion appear to vary more for sentential items, whereas indices for lexical overlap tend to be more highly correlated to intersentential or extra-textual item difficulty. In the case of referential overlap, those indices that did correlate with intersentential difficulty were positive, such that increases in anaphor overlap (CRFANPa) and word variation (CRFCWO1d) reflected increased item difficulty. This makes sense when you consider that texts that contain a high number of references or more variation at the lexical level would result in greater complexity (Eggins, 2004; Trace, 2016). Without local clues to draw upon, a richer understanding of the text at large would be required by examinees in order to answer the items. Curiously, sentential items showed opposite relationships for argument and anaphor overlap at both the level of adjacent sentences (CRFAO1, CRFANP1) and whole passages (CRFANPa). In all cases, as overlap decreased, item difficulty increased. However, as seen below, this was not a significant predictor of item difficulty, so it may be that this was more attributable to random variance than specific patterns in the data.

Semantic overlap showed some relationship with intersentential item difficulty, in that as semantic overlap increased for a passage, item difficulty slightly decreased. This is somewhat in contrast to earlier findings that claim lexical cohesion was more dependent upon locally-available context clues (Bridge & Winograd, 1982; Storey, 1997), however, in both cases this was about the items and not the lexical properties of the passages. Trace (2016) found that lexical cohesion shared only a weak relationship with item difficulty as a whole, so that these indices were slightly more weighted when separated by classification may reflect a more accurate interpretation of the implications of semantic overlap on items, though this relationship remains tenuous.

What is somewhat surprising is that indices for connectives did not reveal a clear pattern for any particular item type with the exception of temporal conjunctions (CNCTemp), which was negatively correlated with all three classifications. Previous research has pointed to conjunctions as having an impact on passage cohesion (Bridge & Winograd, 1982; Eggins, 2004; Trace, 2016), and yet their presence or absence here showed little relationship to item function. The presence of logical connections (CNCLogic) did show a positive relationship with intersentential difficulty, which may indicate that as passage complexity increases, those items relying on a deeper argument structure are also more difficult to answer.

RQ_2: To what degree do measures for cohesion form distinct factors for cloze test passages?
Mostly distinct factors were observed for the subset of indices used here, resulting in components for each *referential cohesion*, *conjunctives*, and *semantic overlap* with minimal complexity. As validation studies already exist for *Coh-Metrix* (see McNamara et al., 2010), the presence of distinct factors was not a surprise on its own. However, given that the dataset contained only 15 unique passages, limited variation should have made observable patterns in the data difficult to observe (Trace, 2016). Because item-level variation was included in these indices (i.e., analyzing the passages with the target word removed), this may have afforded enough variation to capture, at least minimally, relevant differences in passage cohesion similar to what examinees face when answering cloze tests. This provides some validity evidence to

previous notions that cloze passages are made up of different cohesive aspects and furthermore that they impact item difficulty differently (Trace, 2016).

RQ_3: To what degree can cohesion predict cloze item difficulty, and are these predictions similar for items drawing upon different levels of context?

Results of regression analysis point to effects for passage cohesion on intersentential item difficulty, supporting previous findings by Brown and colleagues (2016; Trace et al., 2017; Trace, 2020). These effects remain low (R^2 = .268), however, especially relative to previously reported item-level effects (Trace, 2016). Unlike those same findings, where clear relationships for passage cohesion could not be observed relative to item difficulty, there appears to be a slightly stronger case here for the impact of cohesion on item difficulty, and that even with the minimalizing effects of passage-level quantitative data lacking the kind of variance expected for item-level analyses, differences in predictive power were still observed relative to item type.

It seems that the combination of both word variation (CRFCWO1d) and semantic overlap (LSASSp) played the biggest role in predicting item difficulty, with the two of these indices contributing the majority of the explained variance (~20%), whereas logical connections only accounted for 6% of the variance. This, again, seems to conflict with earlier studies that found that conjunctions were closely related to item difficulty (Bridge & Winograd, 1982; Eggins, 2004; Trace, 2016).

Only three of the possible 23 indices showed any kind of direct, predictive power for item difficulty, meaning that either the other 20 indices were somehow irrelevant, or—more likely—a combined lack of variation between the indices and the relative similarity of what they are measuring means that differences are hard to observe. As with any quantitative measure of passage quality, there remains much that is lost in translation. That being said, as passage-factors were able to partially predict intersentential item difficulty, this would seem to provide further evidence that cloze items can and do draw on passage-level features (Brown et al., 2016; Kleijn et al., 2019; Trace et al., 2017).

Conclusion

So what can be said about cohesion and cloze item function? Unfortunately, while it appears that there is a clear connection between passage factors and item function, the exact nature of this relationship remains difficult to parse in a practical sense. Given that cloze tests already rely on a number of considerations in their very design (i.e., item selection, scoring methods), it should be no surprise that a careful hand is required when both selecting passages and recognizing their effects on interpretations of performance. As JD himself puts it:

> I have discovered that cloze tests by definition involve a sort of tailoring process. Either you will do it on purpose through some sort of rational well-tailored cloze item analysis process or it will occur naturally with tremendous inefficiency because many items will naturally be switched off or discriminating poorly/marginally. (Brown, 2013, p. 26)

In addition to content and grammatical structure, test designers should examine the role of cohesion in potential cloze passages. While that does not necessarily mean putting everything through a rigid text analysis program, having an awareness of a passage's argument structure, use of references, and lexical variation may provide a useful point of comparison when selecting passages, or even provide a basic set of criteria for selecting items that are able to measure both sentential and intersentential comprehension.

What is known is that cloze tests are not a mere set of items in isolation, and that test-designers and researchers should treat them with the same care and caution that other assessments, discrete or otherwise, receive in the name of reliability, validity, and fairness. Their simplicity belies an underlying complexity, as well as a power in that they can provide a somewhat unique perspective on second language ability.

Limitations

The above findings are not without several important limitations, primarily rooted in the very data being used. The use of exact-answer scoring means that interpretations of item performance are limited to whether or not examinees could reproduce the original omitted word from the text, while excluding all other cases where a semantically or syntactically equivalent response may have resulted in a correct answer—and therefore indicate higher command of the language—but were nonetheless marked incorrect due to the scoring constraints. Future research would do well in examining similar kinds of questions using acceptable-answer scoring approaches.

Likewise, passage-level descriptions do not contain equivalent variance to item-level descriptions and so results based upon variance will necessarily be reduced in their explanatory power. While efforts were made to introduce relevant variance into these variables, it still falls short of what is happening qualitatively at the passage level. Lastly, regression models can only go so far in explaining unique variance and specific effects of multiple indices on item difficulty. Ideally, structural models may be a better representation of the impact of not only passage cohesion on item difficulty, but other latent factors as well (e.g., item-factors, syntactic complexity), though these would also require a larger sample size and more clearly observed variances at the passage-level.

In closing[2], it's difficult to overestimate the effect that JD's work has had on our understanding of cloze tests for second language assessment and research purposes over the years. For as much as cloze tests may be looked to as easy or quick solutions to creating and administering tests, their very nature seems to imply layers of complexity that, like any form of assessment, rest upon the diligence and carefulness of the researcher.

ORCID

 https://orcid.org/0000-0001-9010-5946

[2] Perhaps that should be spelled "clozing"?

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal*, *76*(4), 468–479. https://doi.org/10.1111/j.1540-4781.1992.tb05394.x

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219–227. https://doi.org/10.2307/3586211

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556. https://doi.org/10.2307/3586277

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading, 10*, 291–299.

Bridge, C. A., & Winograd, P. N. (1982). Readers' awareness of cohesive relationships during cloze comprehension. *Journal of Reading Behavior*, *14*(3), 299–312. https://doi.org/10.1080/10862968209547457

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal*, *64*(3), 311–317. https://doi.org/10.1111/j.1540-4781.1980.tb05198.x

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 237–250). Newbury House.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83* (pp. 109–119). TESOL.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*, 19–31. https://doi.org/10.1177/026553228800500102

Brown, J. D. (1989). Cloze item difficulty. *JALT Journal*, *11*(1), 46–67.

Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10*, 93–116. https://doi.org/10.1177/026553229301000201

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies*, *7*(1), 1–32.

Brown, J. D., & Grüter, T. (2022). The same cloze for all occasions? *International Review of Applied Linguistics in Language Teaching*, *60*(3), 599–624. https://doi.org/10.1515/iral-2019-0026

Brown, J. D., Janssen, G., Trace, J. & Kozhevnikova, L. (2012). A preliminary study of cloze procedure as a tool for estimating English readability for Russian students. *Second Language Studies*, *31*(1), 1–22. https://doi.org/10.35213/2686-7516-2019-1-1-16-27

Brown, J. D., Trace, J., Janssen, G., & Kozhevnikova, L. (2016). How well do cloze items work and why? In C. Gitsaki & C. Coombe (Eds.), *Current issues in language assessment and evaluation: Research and practice.* Cambridge Scholars.

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor's new cloze: strategies for revising cloze tests. *A focus on language test development: Expanding the language proficiency construct across a variety of tests*, 143–161.

Chihara, T., Oller, J. W., Jr., Weaver, K. A., & Chavez-Oller, M. A. (1994). Are cloze items sensitive to constraints across sentences? In J. W. Oller Jr., & J. Jonz (Eds.), *Cloze and coherence* (pp. 135–148). Associated University Presses. https://doi.org/10.1111/j.1467-1770.1977.tb00292.x

Cooperman, A. W., & Waller, N. G. (2022). Heywood you go away! Examining causes, effects, and treatments for Heywood cases in exploratory factor analysis. *Psychological Methods*, *27*(2), 156–176. https://doi.org/10.1037/met0000384

Dastjerdi, H. V., & Talebinezhad, M. R. (2006). Chain-preserving deletion procedure in cloze: A discoursal perspective. *Language Testing*, *23*(1), 58–72. https://doi.org/10.1191/0265532206lt318oa

Eggins, S. (2004). *An introduction to systemic functional linguistics*. Continuum.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, *31*(1), 16–28. https://doi.org/10.1177/0734282912451971

Goldman, S. R., & Murray, J. D. (1992). Knowledge of connectors as cohesion devices in text: A comparative study of native-English and English-as-a-second-language speakers. *Journal of Educational Psychology*, *84*(4), 504–519. https://doi.org/10.21236/ada213269

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Longman. https://doi.org/10.4324/9781315836010

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*(1), 61–83. https://doi.org/10.2307/3586852

Kleijn, S., Pander Maat, H., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, *36*(4), 553–572. https://doi.org/10.1177/0265532219840382

Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *Modern Language Journal*, *86*(4), 571–586. https://doi.org/10.1111/1540-4781.00162

Linacre, J. (2010). FACETS (Version 3.67.0). MESA Press.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix.* Cambridge University Press. https://doi.org/10.1017/cbo9780511894664.017

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Oller, J. W. Jr., & Jonz, J. (1994). A critical appraisal of related cloze research. In J. W. Oller Jr., & J. Jonz (Eds.), *Cloze and coherence* (pp. 371–408). Associated University Presses

Shanahan, T., Kamil, M., & Tobin, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229–255. https://doi.org/10.2307/747485

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*(2), 214–231. https://doi.org/10.1177/026553229701400205

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Pearson.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 414–438. https://doi.org/10.1177/107769905303000401

Trace, J. (2016). *A validation argument for cloze test item function in second language assessment* (Doctoral dissertation, University of Hawai'i at Manoa).

Trace, J. (2020). Clozing the gap: How far do cloze items measure? *Language Testing*, *37*(2), 235–253. https://doi.org/10.1177/0265532219888617

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, *34*(2), 151–174. https://doi.org/10.1177/0265532215623581

Appendix A

Coh-Metrix Indices (McNamara et al., 2014)

Index	Description
CRFAO1	Argument overlap (i.e., overlap between nouns or pronouns) between adjacent sentences
CRFSO1	Stem overlap (i.e., overlap between nouns and lemmas) between adjacent sentences
CRFNOa	Noun overlap between adjacent sentences
CRFAOa	Argument overlap (i.e., overlap between nouns or pronouns) across the entire passage
CRFCWO1	Content word overlap (proportion) between adjacent sentences
CRFCWO1d	Standard deviation of content word overlap (proportion) between adjacent sentences
CRFCWOad	Standard deviation of content word overlap (proportion) across the entire passage
CRFANP1	Anaphor overlap (i.e., overlap between nouns and pronouns) between adjacent sentences
CRFANPa	Anaphor overlap (i.e., overlap between nouns and pronouns) across the entire passage
LSASS1d	Standard deviation of semantic overlap for adjacent sentence-to-sentence units
LSASSp	Semantic overlap of sentence-to-sentence units across the entire passage
LSASSpd	Standard deviation of semantic overlap of sentence-to-sentence units across the entire passage
LSAPP1	Semantic overlap between adjacent paragraphs
LSAPP1d	Standard deviation of semantic overlap between adjacent paragraphs
LSAGN	Average givenness of each sentence
LSAGNd	Standard deviation of givenness in each sentence
CNCAll	Incidence of all connectives in the passage
CNCCaus	Incidence of all causal connectives in the passage
CNCLogic	Incidence of all logical connectives in the passage
CNCADC	Incidence of all contrastive connectives in the passage
CNCTemp	Incidence of all temporal connectives in the passage
CNCAdd	Incidence of all causal connectives in the passage
CNCPos	Incidence of all positive connectives in the passage